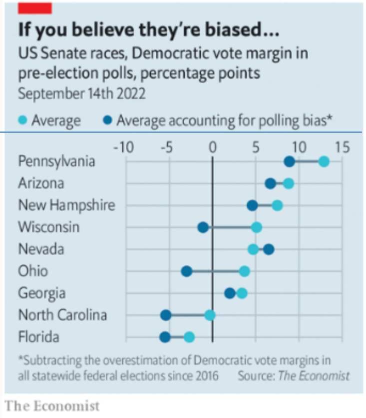
Methods of Applied Statistics I

STA2101H F LEC9101

Week 5

October 12 2022



Today

Start Recording

1. Upcoming events
2. Recap
3. Steps in analysis; types of studies
4. In the News

5. HW 4 3rd hour

6. Sections for Project
    - a description of the scientific problem of interest
    - how (and why) the data being analyzed was collected
    - preliminary description of the data (plots and tables)
    - models and analysis
    - summary for a statistician of the analysis and conclusions
    - non-technical summary for a non-statistician of the analysis and conclusions

Upcoming

- October 13 3.30-4.30 : DoSS Seminar Room 9014 (Hydro Building)
- Brenda Betancourt, U Chicago
- Microclustering for record linkage applications

Brenda Betancourt
NORC – University of Chicago

Brenda is currently a Senior Statistician at NORC at the University of Chicago. Before joining NORC, she was an Assistant Professor in the Department of Statistics at the University of Florida. She obtained her PhD in Statistics at the University of California, Santa Cruz and was a postdoctoral fellow at Duke University working on Bayesian models and algorithms for record linkage and Network analysis.

Microclustering for record linkage applications

Upcoming

- October 15 12.00-1.00 Toronto Data Workshop; Room BL520, Bissell Building and online link
  data_4_lyf
- April Wang, U Michigan

  "Reimagining Tools for Collaborative Data Science"

Recap

- Model Selection: hierarchical principle, testing procedures ("$p < 0.05$")
- criterion-based procedures (AIC, BIC, C_p, R_a^2)
- regularization/penalization methods: Lasso

- Model Building: plots, partial plots
- consideration of units and types of variation
- potential transformation of variables HW3

- clarification of objecives: prediction and/or explanation
- criterion-based methods may be helpful for prediction
- automated methods rarely useful for explanation
- there may be several models consistent with the data

Design of Studies

- common objectives
- to avoid systematic error, that is distortion in the conclusions arising from sources that do not cancel out in the long run
- to reduce the non-systematic (random) error to a reasonable level by replication and other techniques
- to estimate realistically the likely uncertainty in the final conclusions
- to ensure that the scale of effort is appropriate

… design of studies

- we concentrate largely on the careful analysis of individual studies
- in most situations synthesis of information from different investigations is needed

- but even there the quality of individual studies remains important
- examples include overviews (such as the Cochrane reviews)

- in some areas new investigations can be set up and completed relatively quickly; design of individual studies may then be less important

… design of studies

- formulation of a plan of analysis
- establish and document that proposed data are capable of addressing the research questions of concern

- main configurations of answers likely to be obtained should be set out
- level of detail depends on the context

- even if pre-specified methods must be used, it is crucial not to limit analysis
- planned analysis may be technically inappropriate
- more controversially, data may suggest new research questions or replacement of objectives

- latter will require confirmatory studies

Unit of study and analysis

- smallest subdivision of experimental material that may be assigned to a treatment
  context: Expt
- Example: RCT – unit may be a patient, or a patient-month (in crossover trial)
- Example: public health intervention – unit is often a community/school/...

- split plot experiments have two classes of units of study and analysis
- in investigations that are not randomized, it may be helpful to consider what the primary unit of analysis would have been, had a randomized experiment been feasible

- the unit of analysis may not be the unit of interpretation – ecological bias
  systematic difference between impact of x at different levels of aggregation

Ecological bias

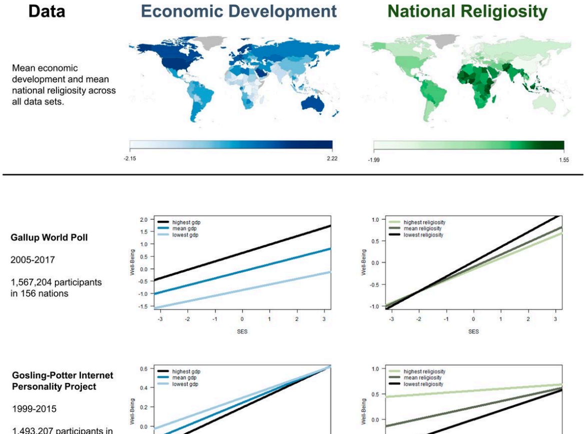
- CD: Illustration – “For country- or region-based mortality data, countries of regions respectively may reasonably constitute the units of analysis with which to assess the relationship of the data to dietary and other features
- “Yet the objective is interpretation at an individual person level
- “The situation may be eased if supplementary data on explanatory variables are available at the individual level, because this may clarify the connection of between-unit and within-unit variation”

Example

Berkessel et al., PNAS, 2021

- feelings of well-being are associated with socio-economic status link
- the strength of the association is larger in developed nations than in developing nations
- conventional explanation: in nations with a high level of economic development perhaps higher SES carries some intrinsic value
- this paper: in nations with a high level of religiosity, the strength of the association between SES and well-being is weaker
- religiosity could attenuate the link between well being and SES Economist, Sep 25 2021

Distribution of national economic development, national religiosity, and estimated means of the cross-level interactions (model 3).



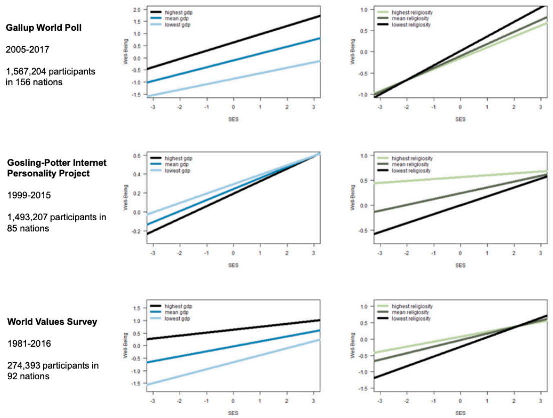


Fig. 2. Distribution of national economic development, national religiosity, and estimated means of the cross-level interactions (model 3). The top row depicts mean national economic development and mean national religiosity worldwide (z-standardized and averaged across data sets). Lighter colors represent lower values, darker colors higher values. The three bottom rows depict estimated marginal means of the cross-level interactions when both national moderators were included in the model (i.e., model 3). Depicted are the moderating effects of national economic development and national religiosity on the association between SES and well-being in all three data sets.

Depicted are the moderating effects of national economic development and national religiosity on the association between SES and well-being in all three data sets.

ANCOVA in PNAS paper

see also LM-2, 14.1; LM-1, 13.1

- $\mathrm{E}(y_i \mid x_i, z_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i$ `WellB ~ ses + relig + ses:relig`
- y_i: Well-being, x_i: Socio-economic status, z_i: Religiosity – a factor variable
- z_i = (“low”, “medium”, “high”) `model.matrix`

$$\begin{aligned}
\mathrm{E}(y_i \mid x_i, z_i = \text{"}low\text{"}) &= \beta_0 + \beta_1 x_i \\
\mathrm{E}(y_i \mid x_i, z_i = \text{"}med\text{"}) &= \beta_0 + \beta_2 + (\beta_1 + \beta_4) x_i \\
\mathrm{E}(y_i \mid x_i, z_i = \text{"}hi\text{"}) &= \beta_0 + \beta_3 + (\beta_1 + \beta_5) x_i
\end{aligned}$$

- as usual, the paper’s a bit more complicated
- some data collected on people, and some on countries – multi-level model
- “Following standard practice, we averaged person-level religiosity within each nation ”
- there’s another covariate – GDP
- “Following a standard economic method, we log-transformed the GDP data”

Factor variables: modelling

- a factor variable is treated as categorical
- a non-factor variable is treated as continuous
- it depends on the application which is preferred

- a linear model with one factor and one continuous variable might be written as, for example:
$$y_{ij} = \mu + \alpha_j + \beta x_{ij} + \epsilon_{ij}, \quad j = 1, \dots, J; \quad i = 1, \dots m$$
- linear in x, but arbitrary changes in $\mathbb{E}(y)$ by category (here indexed by j)
- `R` doesn't distinguish this at the modelling phase:
`lm(response ~ variable1 + variable2, data = ...)`
- but uses metadata in the data frame to accommodate factors
- `is.factor(variable)` and `newvar <- as.factor(oldvar)` are helpful

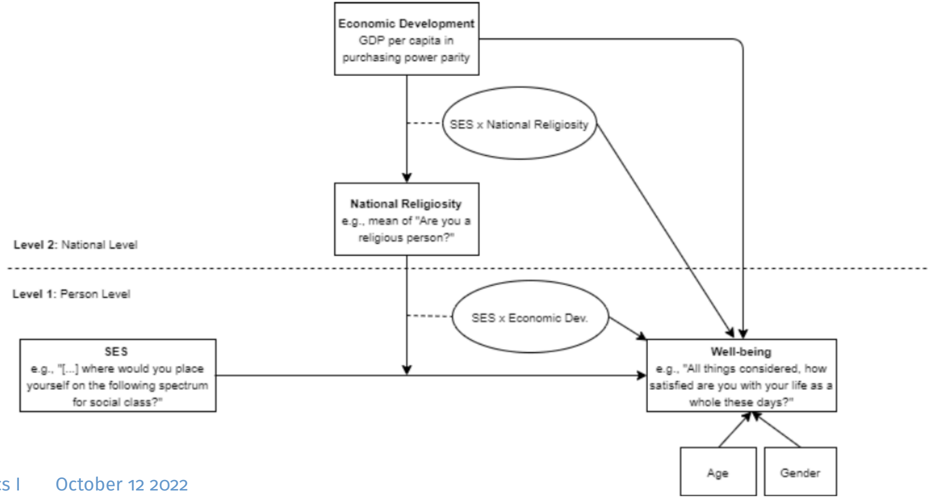


Figure S1. Mixed Effects Mediated Moderation Model (model 4). Statistical model to calculate the portion of the cross-level interaction

Design of studies: randomized experiments

- unit of analysis – “smallest subdivision of the experimental material such that two distinct units might be randomized to different treatments”
  - example: patient in a clinical trial
  - example: plot of land in an agricultural trial
  - example: units of material in a quality control trial
- advantages of randomization?
  - balances other potential influences on responses
  - avoidance of systematic error
  - a systematic difference in response not due to treatment under study

Avoidance of systematic error

- "distortion in the conclusions arising from irrelevant sources that do not cancel out in the long run"
- can arise through systematic aspects of, for example, a measuring process, or the spatial or temporal arrangement of units
- this can often be avoided by design, or adjustment in analysis
- can arise by the entry of personal judgement into some aspect of the data collection process
- this can often be avoided by randomization and blinding

- “treatment” is not assigned to units, only observed
- any observed effect of treatment cannot be assumed to be causal

“correlation is not causation”

- we can try to assess the effect by controlling for other variables that may also influence the response
- but we cannot control for unmeasured variables

418 9 · *Designed Experiments*

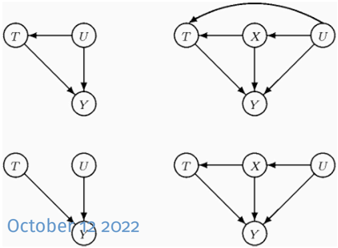


Figure 9.1 Directed acyclic graphs showing consequences of randomization. An arrow from T to Y indicates dependence of Y on T, and so forth. In general both response Y and treatment T may depend on properties U of units (upper left). Randomization (lower left) makes treatments and units independent, so any observed dependence of Y on T cannot be ascribed to joint dependence on U. The upper right graph shows the general dependence of Y, T, and covariates X on U.

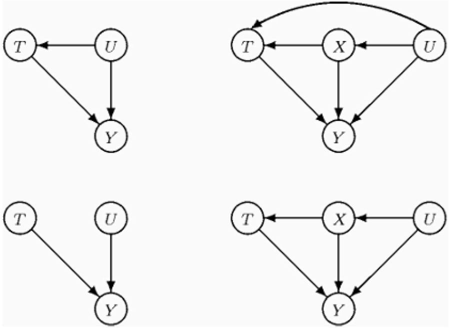


Figure 9.1 Directed acyclic graphs showing consequences of randomization. An arrow from T to Y indicates dependence of Y on T, and so forth. In general both response Y and treatment T may depend on properties U of units (upper left). Randomization (lower left) makes treatments and units independent, so any observed dependence of Y on T cannot be ascribed to joint dependence on U. The upper right graph shows the general dependence of Y, T, and covariates X on U. Randomization makes T and U independent, conditional on X (lower right), so any influence of U on T is mediated through X, for which adjustment is possible in principle. Thus having adjusted for X, dependence of Y on T cannot be due to U.

the control group. The response is to be the blood pressure of an individual measured a fixed time after the drug has first been administered. We calculate the average changes for the treated and control groups, $\overline{y}_1$ and $\overline{y}_0$, observe that $\overline{y}_1 - \overline{y}_0$ is significantly less than zero, and declare that the drug plays an effect in reducing blood pressure. Is this headline news? No!

A key difficulty is that the procedure does not avoid biased allocation of treatments to units. For example, if the control group mostly consisted of those patients with

Support for causation

- strength of the association
- consistency of the association
- specificity of the proposed causal factor
- potential cause occurs before its effect (temporality)
- dose-response relationship
- a subject-matter theory exists

- “natural experiments” e.g. minimum wage

Types of observational studies

- secondary analysis of data collected for another purpose
- estimation of some feature of a defined population

could in principle be found exactly

- tracking across time of such features
- study of a relationship between features, where individuals may be examined
  - at a single time point
  - at several time points for different individuals
  - at different time points for the same individual
- census
- meta-analysis: statistical assessment of a collection of studies on the same topic

Binary responses

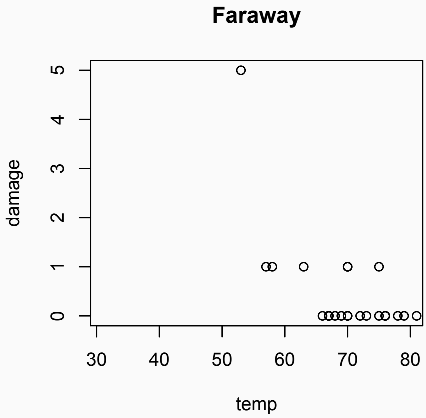
- simple linear regression $E(y_i \mid x_i) = \beta_0 + \beta_1 x_i, \quad \text{var}(y_i \mid x_i) = \sigma^2$
- suppose $y \in \{0, 1\}$
- examples
- $E(y_i \mid x_i) =$

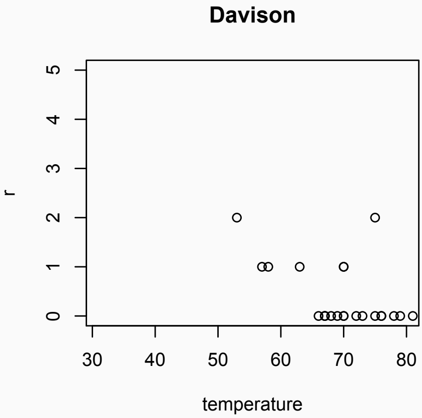
Binomial Data

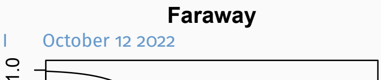
FELM, Ch. 2, SM, Ch.1; §4.4,5;

Table 1.3 O-ring thermal distress data. r is the number of field-joint O-rings showing thermal distress out of 6, for a launch at the given temperature (°F) and pressure (pounds per square inch) (Dalal *et al.*, 1989).

Flight	Date	Number of O-rings with thermal distress, r	Temperature (°F) x_1	Pressure (psi) x_2
1	21/4/81	0	66	50
2	12/11/81	1	70	50
3	22/3/82	0	69	50
5	11/11/82	0	68	50
6	4/4/83	0	67	50
7	18/6/83	0	72	50
8	30/8/83	0	73	100
9	28/11/83	0	70	100
41-B	3/2/84	1	57	200
41-C	6/4/84	1	63	200
41-D	30/8/84	1	70	200
41-G	5/10/84	0	78	200
51-A	8/11/84	0	67	200
51-C	24/1/85	2	53	200
51-D	12/4/85	0	67	200
51-B	29/4/85	0	75	200
51-G	17/6/85	0	70	200
51-F	29/7/85	0	81	200
51-I	27/8/85	0	76	200
51-J	3/10/85	0	79	200
61-A	30/10/85	2	75	200
61-B	26/11/86	0	76	200
61-C	21/1/86	1	58	200


Faraway
damage
0
1
2
3
4
5
30
40
50
60
70
80
temp


Davison
r
0
1
2
3
4
5
30
40
50
60
70
80
temperature


Faraway
1.0


Davison
1.0

Table 1. O-Ring Thermal-Distress Data

		Field			Nozzle				Leak-check pressure	
Flight	Date	Erosion	Blowby	Erosion or blowby	Erosion	Blowby	Erosion or blowby	Joint temperature	Field	Nozzle
1	4/12/81							66	50	50
2	11/12/81	1		1				70	50	50
3	3/22/82							69	50	50
5	11/11/82							68	50	50
6	4/04/83				2		2	67	50	50
7	6/18/83							72	50	50
8	8/30/83							73	100	50
9	11/28/83							70	100	100
41-B	2/03/84	1		1	1		1	57	200	100
41-C	4/06/84	1		1	1		1	63	200	100
41-D	8/30/84	1		1	1	1	1	70	200	100
41-G	10/05/84							78	200	100
51-A	11/08/84							67	200	100
51-C	1/24/85	2, 1*	2	2		2	2	53	200	100
51-D	4/12/85				2		2	67	200	200
51-B	4/29/85				2, 1*	1	2	75	200	100
51-G	6/17/85				2	2	2	70	200	200
51-F	7/29/85				1			81	200	200
51-I	8/27/85				1			76	200	200
51-J	10/03/85							79	200	200
61-A	10/30/85		2	2	1			75	200	200
61-B	11/26/85				2	1	2	76	200	200
61-C	1/12/86	1		1	1	1	2	58	200	200
61-I	1/28/86							31	200	200
	Total	7, 1*	4	9	17, 1*	8	17			

*Secondary O-ring.

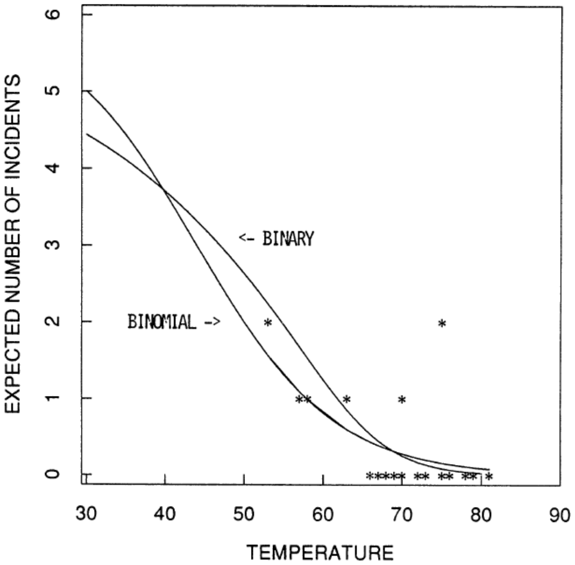


Figure 4. O-Ring Thermal-Distress Data: Field-Joint Primary O-Rings, Binomial-Logit Model, and Binary-Logit Model.

Modelling numbers/proportions of events

- $y_i \sim Bin(6, p_i), \quad i = 1, \ldots, 23$
- in general: n_i trials, y_i successes, probability of success p_i
- for regression: associated covariate vector x_i, e.g. temperature
- SM uses m_i and r_i instead of n_i and y_i
- each y_i could in principle be the sum of n_i independent Bernoulli trials
- each of the n_i trials having the same probability p_i
- with the same covariate vector x_i

FELM 'covariate classes', p.26

Challenger data: Faraway

```
> library(faraway); data(orings)
> logitmod <- glm(cbind(damage,6-damage) ~ temp, family = binomial, data = orings)
> summary(logitmod)
Call:
glm(formula = cbind(damage, 6 - damage) ~ temp, family = binomial,
    data = orings)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.66299    3.29626   3.538 0.000403 ***
temp        -0.21623    0.05318  -4.066 4.78e-05 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38.898  on 22  degrees of freedom
Residual deviance: 16.912  on 21  degrees of freedom
```

Challenger data: Davison

```
> library(SMPracticals) # this is for datasets in
                        #Statistical Models by Davison
> data(shuttle) # same example, different name
> shuttle2 <- data.frame(as.matrix(shuttle)) # this is a kludge to avoid
                                 #an error with head(shuttle)
> head(shuttle2)
  m r temperature pressure
1 6 0          66       50
2 6 1          70       50
3 6 0          69       50
4 6 0          68       50
5 6 0          67       50
6 6 0          72       50
> par(mfrow=c(2,2)) # puts 4 plots on a page


> with(orings,plot(temp,damage,main="Faraway",xlim=c(31,80)))
> with(shuttle,plot(temperature,r,main="Davison",xlim=c(31,80),
+ ylim=c(0,5)))
```

Challenger data fits